Detecting semantic shifts in Slovene Twitterese

Darja Fišer[1,2] Nikola Ljubešić[2,3]

[1]Department of Translation
Faculty of Arts
University of Ljubljana

[2]Department of Knowledge Technologies
Jožef Stefan Institute

[3]Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences
University of Zagreb

RASLAN, 2 Dec 2016

Problem

- meanings of words are not fixed but undergo changes
  - advent of new senses
  - established senses take new shades of meaning or become obsolete
- semantic shifts typically occur systematically
  - expand (`miška/mouse`)
  - narrow down (`faks/fax`)
  - amelioration (`hudo/terrific`)
  - pejoration (`blondinka/blond woman`)

Method

- learn word embeddings to model meaning of words
- learn separate word representations for non-standard and standard corpus occurrences
- semantic shift (ss) of a word (w) as cosine distance between dense representation of the word learned from the standard (w_s) and nonstandard corpus (w_n).

$$ss(w) = 1 - cossim(\vec{w_s}, \vec{w_n})$$

- calculate semantic shift $ss(w)$ over the whole lexicon (headwords) and manually inspect those with highest ss

Data

Non-standard corpus

- 100-million token corpus of Slovene tweets (Fišer et al. 2016)

Reference corpus

- 1-billion token reference corpus Gigafida (Logar et al. 2012)

Data

Non-standard corpus

- 100-million token corpus of Slovene tweets (Fišer et al. 2016)

Reference corpus

- 1-billion token reference corpus Gigafida (Logar et al. 2012)

Headwords and features

- headwords: lowercased lemmata extended with first two characters of morphosyntactic description (`miška#Nc`)
- filtered by Sloleks lexicon, occurring more than 500 times in the non-standard dataset – 5425 headwords
- features: surface forms in a window of ± 2

Implementation

Requirement 1

- need different representation of headwords (`miška#Nc`) and features (`tuki`)
- word2vec, GloVe, fastText do not meet the requirements
- word2vecf enables preparing (*headword*, *feature*) pairs in advance

Implementation

Requirement 1

- need different representation of headwords (`miška#Nc`) and features (`tuki`)
- word2vec, GloVe, fastText do not meet the requirements
- word2vecf enables preparing (*headword*, *feature*) pairs in advance

Requirement 2

- need to learn representations from both corpora simultaneously
- apply a simple trick: corpus the headword occurred in as a prefix (`n_miška#Nc`)

Linguistic analysis

Selection criterion

- top-ranking 200 lemmas from reference & Twitter corpus which display the most differences in contexts

Analysis procedure

- preprocessing errors (`talka` - `talk`)
- remaining 110 lemmas compared through Word Sketches
  - no semantic shift
  - minor semantic shift
    - semantic narrowing (`posodobiti`, `podnapis`)
    - different usage pattern (`kvadrat`, `eter`)
    - redistribution of senses (`odklop`, `sesalec`)
  - major semantic shift
    - CMC-specific (`sledilec`, `opomnik`)
    - colloquial (`optika`, `carski`)
    - event-related (`vztrajnik`, `pirat`)

Results

	No.	%
No shift	28	25%
Minor shift	21	19%
Semantic narrowing	3	3%
Usage pattern	6	5%
Redistribution of senses	12	11%
Major shift	61	56%
CMC-specific	6	5%
Colloquial	23	21%
Events	32	29%

Discussion

- some type of semantic shift detected in 75% cases (quite accurate, given the complexity of the task)
- 74% of all the shifts detected were major (quite suitable, given the task)
- 50% shifts due to daily events & informal language (most interesting cases, missing in dictionaries)
- some highly creative attribution of new meaning to common words (`kahla/potty` - politician Karel Erjavec)
- minor shifts systematically show the differences between the two corpora (many more novel usages than narrowings - the reference corpus could be enhanced with social media texts)

Conclusion

Summary

- first results of semantic shift detection for Slovene
- distance of WE representation in reference & Twitter corpus
- many valuable semantic shift candidates, esp. due to daily events & informal communication
- used for dictionary of Twitterese

Future work

- extend manual analysis to lower-ranked candidates
- extend approach to lower-frequency candidates
- compare with alternative methods, e.g. representing words as word sketches / syntactic patterns
- use supervised learning, discriminating between types of semantic shifts or filtering preprocessing errors

Detecting semantic shifts in Slovene Twitterese

Darja Fišer[1,2] Nikola Ljubešić[2,3]

[1]Department of Translation
Faculty of Arts
University of Ljubljana

[2]Department of Knowledge Technologies
Jožef Stefan Institute

[3]Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences
University of Zagreb

RASLAN, 2 Dec 2016